Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2012, **3**, 116–128 doi: 10.1111/j.2041-210X.2011.00124.x

Applying additive modelling and gradient boosting to assess the effects of watershed and reach characteristics on riverine assemblages

Kelly O. Maloney[1,2]*†, Matthias Schmid[3]† and Donald E. Weller[1]

[1]*Smithsonian Environmental Research Center, 647 Contees Wharf Road, PO Box 28, Edgewater, MD 21037-0028, USA;* [2]*USGS, Leetown Science Center, Northern Appalachian Research Laboratory, 176 Straight Run Road, Wellsboro, PA 16901, USA; and* [3]*Department of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander-University Erlangen-Nuremberg, Waldstrasse 6, 91054 Erlangen, Germany*

Summary

1. Issues with ecological data (e.g. non-normality of errors, nonlinear relationships and autocorrelation of variables) and modelling (e.g. overfitting, variable selection and prediction) complicate regression analyses in ecology. Flexible models, such as generalized additive models (GAMs), can address data issues, and machine learning techniques (e.g. gradient boosting) can help resolve modelling issues. Gradient boosted GAMs do both. Here, we illustrate the advantages of this technique using data on benthic macroinvertebrates and fish from 1573 small streams in Maryland, USA.

2. We assembled a predictor matrix of 15 watershed attributes (e.g. ecoregion and land use), 15 stream attributes (e.g. width and habitat quality) and location (latitude and longitude). We built boosted and conventionally estimated GAMs for macroinvertebrate richness and for the relative abundances of macroinvertebrates in the Orders Ephemeroptera, Plecoptera and Trichoptera (%EPT); individuals that cling to substrate (%Clingers); and individuals in the collector/gatherer functional feeding group (%Collectors). For fish, models were constructed for taxonomic richness, benthic species richness, biomass and the relative abundance of tolerant individuals (%Tolerant Fish).

3. For several of the responses, boosted GAMs had lower pseudo R^2s than conventional GAMs for in-sample data but larger pseudo R^2s for out-of-bootstrap data, suggesting boosted GAMs do not overfit the data and have higher prediction accuracy than conventional GAMs. The models explained most variation in fish richness (pseudo $R^2 = 0{\cdot}97$), least variation in %Clingers (pseudo $R^2 = 0{\cdot}28$) and intermediate amounts of variation in the other responses (pseudo R^2s between 0·41 and 0·60). Many relationships of macroinvertebrate responses to anthropogenic measures and natural watershed attributes were nonlinear. Fish responses were related to system size and local habitat quality.

4. For impervious surface, models predicted below model-average macroinvertebrate richness at levels above *c.* 3·0%, lower %EPT above *c.* 1·5%, and lower %Clingers for levels above *c.* 2·0%. Impervious surface did not affect %Collectors or any fish response. Prediction functions for %EPT and fish richness increased linearly with $\log_{10}$ (watershed area), %Tolerant Fish decreased with $\log_{10}$ (watershed area), and benthic fish richness and biomass both increased nonlinearly with $\log_{10}$ (watershed area).

5. Gradient boosting optimizes the predictive accuracy of GAMs while preserving the structure of conventional GAMs, so that predictor–response relationships are more interpretable than with other machine learning methods. Boosting also avoids overfitting the data (by shrinking effect estimates towards zero and by performing variable selection), thus avoiding spurious predictor effects

*Correspondence author. E-mail: kmaloney@usgs.gov
†Contributed equally to the contents of this manuscript.
Correspondence site: http://www.respond2articles.com/MEE/

and interpretations. Thus, in many ecological settings, it may be reasonable to use boosting instead of conventional GAMs.

Key-words: benthic macroinvertebrates, diversity, fish, generalized additive models, richness, spatial autocorrelation, streams

Introduction

Ecologists often use regression to elucidate relationships among variables and to build predictive models, but issues associated with ecological data and modelling complicate regression analyses. Ecological data are complex, often including non-normal errors, nonlinear relationships and variables that are spatially or temporally autocorrelated. To address these complexities, ecologists routinely apply flexible modelling approaches. For example, generalized linear models (GLMs; McCullagh & Nelder 1989) allow users to specify appropriate response distributions with link functions and to pre-specify nonlinear relationships, such as logarithmic transformations for positive predictor variables. However, pre-specification of nonlinearities is often intractable when relationships are unknown or when the number of relationships is large. Ecologists have applied generalized additive models (GAMs; Yuan & Norton 2003; Austin 2007) to overcome the limitations of GLMs. GAMs do not require pre-defined specification of nonlinearities, but preserve the ability of GLMs to construct complex models (Hastie & Tibshirani 1990; Hastie, Tibshirani & Friedman 2009). In addition, GAMs automatically identify nonlinearities using flexible nonlinear modelling approaches (usually based on spline smoothing) and preserve the easy interpretability of predictor–response relationships of GLMs (Hastie & Tibshirani 1990; Wood 2006).

Ecologists should also consider general modelling issues, like overfitting, variable selection and prediction. Overfitting often results from including too many covariates for a given sample size and yields overly complex models that contain spurious effects. Overfitting also decreases prediction accuracy (Hastie, Tibshirani & Friedman 2009). Variable selection is the process of correctly identifying the subset of covariates that are most important in explaining variation in the response and excluding covariates that do not add explanatory value to a model. Methods available to address overfitting and variable selection include penalized estimation (e.g. the lasso or ridge regression), cross-validation, pruning of decision trees, early stopping of boosting algorithms, model selection using criteria like AIC (see Hastie, Tibshirani & Friedman 2009), and Bayesian regularization (O'Hara & Sillanpää 2009).

Prediction accuracy is an especially important requirement of ecological models, and ecologists have applied machine learning algorithms (e.g. bagging, boosting, random forests, Breiman 1996, 2001; Freund & Schapire 1996) to increase prediction accuracy over standard regression methods (e.g. Cutler *et al.* 2007; Elith, Leathwick & Hastie 2008; Maloney *et al.* 2009). Machine learning procedures also incorporate methods to address overfitting and model selection. Unfortunately, some machine learning techniques (bagging or random forests) produce estimates of predictor–response relationships (marginal functions) that are difficult to interpret because they are based on complex ensembles of decision trees (Cutler *et al.* 2007). Ecologists need modelling approaches that combine the increased prediction accuracy of machine learning algorithms with the interpretability and flexibility of GAM models.

Here, we present a recently developed technique that extends the procedure of gradient boosting to GAMs (Bühlmann & Hothorn 2007), hereafter referred to as boosted GAMs. To illustrate the advantages of this method, we develop boosted GAM models that identify the relationships between watershed-scale environmental and anthropogenic factors and eight measures of small-stream communities within Maryland, USA. All models account for spatial dependencies in the data. We use bootstrapping to compare the prediction accuracies of traditional and boosted GAMs. Although our example focuses on stream data, it demonstrates how boosted GAMs can be used for modelling basic and applied ecological questions in other systems.

Generalized additive models

The model equation of a GAM is formally given by

$$g[\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})] = f_{GAM}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p),$$

where $\boldsymbol{Y}$ denotes a response variable; $\boldsymbol{X}$ is a matrix containing vectors of observations of p explanatory variables; $\boldsymbol{X}_1,\ldots, \boldsymbol{X}_p$ is a set of predictor variables; f_{GAM} is an additive prediction function of the predictor variables; and g is a pre-specified link function relating the conditional mean of $\boldsymbol{Y}$ to the prediction function. For example, if $\boldsymbol{Y}$ is a count response following a Poisson distribution, g will typically be the logarithmic transformation. For simplicity, we consider only the main-effects specification of GAMs, in which f_{GAM} is the sum of a constant intercept term (β_0) and p unknown marginal prediction functions $f_1(\boldsymbol{X}_1),\ldots, f_p(\boldsymbol{X}_p)$, where each marginal function depends on one predictor. Thus, estimating f_{GAM} requires estimating β_0 and $f_1(\boldsymbol{X}_1),\ldots, f_p(\boldsymbol{X}_p)$ given by

$$f_{\mathrm{GAM}}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p) = \beta_0 + \sum_{j=1}^{p} f_j(\boldsymbol{X}_j).$$

The shapes of marginal prediction functions depend on the scales of the predictors. For continuous predictors, smooth (i.e. continuous and differentiable) functions $f_j(\boldsymbol{X}_j)$ that are either linear or nonlinear in $\boldsymbol{X}_j$ are usually tested for inclusion into the GAM. For categorical predictors, dummy-coded

variables are typically used. To ensure that marginal functions are identifiable, $f_1(\boldsymbol{X}_1),\ldots, f_p(\boldsymbol{X}_p)$ are restricted to have zero mean. For a more detailed description of GAMs, see Hastie & Tibshirani (1990).

Gradient boosting

We used gradient boosting to increase predictive ability and to solve two common problems in implementing GAMs: variable selection (choosing the most informative subset of covariates) and model choice (selecting the best representations of continuous predictor variables, e.g. nonlinear vs. linear effects). Boosting is a generic fitting procedure for parametric and nonparametric statistical models. It is one of the most important regression fitting techniques when overfitting and variable selection need to be addressed and is increasingly used in ecology (De'ath 2007; Elith, Leathwick & Hastie 2008; Hastie, Tibshirani & Friedman 2009). Originally, boosting was designed as a machine learning procedure for improving prediction of binary outcomes using weighted ensembles of decision trees (called base-learners) (Freund & Schapire 1996, 1997). Subsequent articles showed that boosting can be regarded as a gradient-descent algorithm in function space and can be used to fit statistical regression models (gradient boosting, Breiman 1998, 1999; Friedman, Hastie & Tibshirani 2000; Friedman 2001).

Gradient boosting is an iterative process. It begins with an initial estimation of a function using a constant offset that fits the data poorly. The fit is improved in each successive iteration by fitting a base-learner (e.g. a tree or least squares estimator) to the negative gradient of a pre-specified loss function (for example, the negative log-likelihood of a GAM). The estimate of the prediction function is updated with the estimate of the negative gradient, and the function approaches the true form of the relationship with successive iterations (see Fig. 1 and the more detailed description below).

Gradient boosting improves model accuracy while simultaneously accomplishing variable selection and model choice, and it has distinct advantages over alternative methods. If gradient boosting is stopped before convergence, it improves prediction accuracy by shrinking regression coefficients towards zero, a method analogous to lasso regression (Tibshirani 1996), ridge regression (Hoerl & Kennard 1970) and shrinkage smoothing (Wood 2006). To accomplish variable selection, gradient boosting sets some coefficients to zero (similar to lasso regression). Wood's shrinkage smoothers accomplish variable selection by imposing heavy penalties on some of the effects, shrinking them to values that are very close to zero. Ridge regression, in contrast, does not accomplish variable selection because all coefficient estimates differ from zero. An additional strength of gradient boosting is its greater flexibility to incorporate nonlinear relationships and spatial effects than lasso regression. Recent work has combined lasso estimation with modelling techniques that incorporate nonlinear relationships (Meier, van de Geer & Bühlmann 2009), but this method has not been adapted yet to regression models with two-dimensional spatial effects or scale parameter estimation.

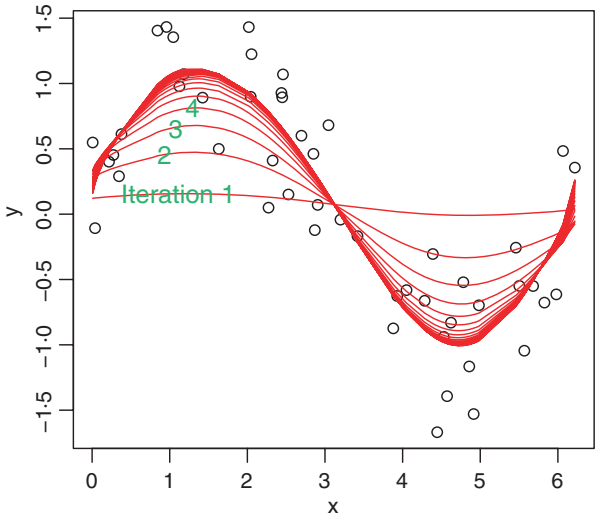

Fig. 1. Marginal function estimates from a simulated data set as the number of boosting iterations increases. The true relationship is $Y = \sin(X) + e$, where e is a normally distributed error with zero mean. Gradient boosting with the squared error loss was used to fit the predictor–response relationship (marginal function). A P-spline base-learner was used for X, v was set equal to 0·1, and the initial boosting estimate (offset value) was set equal to 0 (this corresponds to a horizontal line at 0 – see section *Formal definition of gradient boosting* for a definition of these terms). As the algorithm proceeds, the iteration number increases and the function estimates approach the sine function.

Many boosting algorithms for regression models have been suggested; for example, Cutler *et al.* (2007) and Elith, Leathwick & Hastie (2008) recommended boosting with regression tree base-learners for ecological applications. Also, many types of base-learners (e.g. smoothing or P-splines to fit smooth functions of predictors) have been suggested (see Bühlmann & Yu 2003; Bühlmann & Hothorn 2007; Kneib, Hothorn & Tutz 2009), and gradient boosting was explicitly used to fit GAMs (Bühlmann & Hothorn 2007; Kneib, Hothorn & Tutz 2009). Here, we use gradient boosting with component-wise base-learners, a modification particularly suited for shrinkage and variable selection (Bühlmann & Hothorn 2007).

FORMAL DEFINITION OF GRADIENT BOOSTING

Consider a set of realizations of the response variable $\boldsymbol{Y}$ and the vector of predictor variables $\mathbf{X} := (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$. In this article, the realizations (denoted by $(X_1, Y_1), \ldots, (X_n, Y_n)$) contain data from a published stream assessment, and n denotes the number of sample sites. Define $X := (X_1, \ldots, X_n)$ and $Y := (Y_1, \ldots, Y_n)$. Gradient boosting estimates the optimal prediction function

$$f^* := \arg\min_f E_{Y,X}[\rho(\boldsymbol{Y}, f(\boldsymbol{X}))],$$

i.e., f^* minimizes the expectation of a loss function ρ (here, the negative log-likelihood function of the GAM) over the set of all possible prediction functions f that take

the predictors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$ as input variables. f^* could be any type of function that minimizes $E_{Y,X}[\rho(\boldsymbol{Y}, f(\boldsymbol{X}))]$, but the additive structure of a GAM emerges from the gradient boosting algorithm (below).

The exact distributions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ (needed to derive $E_{Y,X}[\rho(\boldsymbol{Y}, f(\boldsymbol{X}))]$) are typically unknown, so gradient boosting instead minimizes the *empirical risk*

$$R := \frac{1}{n}\sum_{i=1}^{n} \rho(Y_i, f(X_i))$$

over f, where $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are some (independent and identically distributed) sample points of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. R is an approximation of the unknown theoretical risk $E_{Y,X}[\rho(\boldsymbol{Y}, f(\boldsymbol{X}))]$, so R corresponds to the empirical mean of the negative log-likelihood contributions $\rho(Y_i, f(X_i))$, $i = 1, \ldots, n$, of the sample sites with observed data values $(X_1, Y_1), \ldots, (X_n, Y_n)$ substituted into ρ. Minimizing R over f is equivalent to *maximizing* the empirical log-likelihood function R over f, but the former approach is commonly used (Bühlmann & Hothorn 2007).

Gradient boosting uses gradient descent to estimate the set of values $\hat{f}_1 := \hat{f}(X_1), \ldots, \hat{f}_n := \hat{f}(X_n)$ that minimize R over $f_1 := f(X_1), \ldots, f_n := f(X_n)$. Gradient descent is a standard method for function minimization (Hastie, Tibshirani & Friedman 2009). It begins with arbitrary raw estimates (e.g. all zero) of the predicted values $\hat{f}_1, \ldots, \hat{f}_n$, and then iteratively updates the function estimates by adding at each step a small fraction of the negative derivative of R with respect to $f_1, \ldots, f_n$ (also called the negative gradient), evaluated at the current estimates. In this way, the procedure effectively descends the empirical risk surface along the direction with steepest gradient until the minimum of R is reached. The procedure is terminated when it converges, that is, when the current function estimates differ less than a small, user-defined amount from their values in the previous iteration.

If gradient descent is used without considering predictor variables (i.e. if $\hat{f}_1, \ldots, \hat{f}_n$ are treated as fixed parameters independent of $X_1, \ldots, X_n$), then gradient descent yields estimates $\hat{f}_1, \ldots, \hat{f}_n$ that equal observed response values $Y_1, \ldots, Y_n$ because R becomes minimal if $\left(\hat{f}_1, \ldots, \hat{f}_n\right) = (Y_1, \ldots, Y_n)$. Such function estimates would overfit the data and could not make predictions of novel values of $\mathbf{Y}$. Gradient boosting addresses this problem by replacing the negative gradient of R with respect to $f_1, \ldots, f_n$ (denoted by U) with an *estimate* of this gradient that depends on the values of one or several of the predictors. In each iteration of the algorithm, the estimate of U is obtained by evaluating a set of *base-learners*. A base-learner is a regression estimator with U as the outcome variable and a subset of the predictors as input. For example, each base-learner could depend on exactly one of p predictors, yielding a set of p base-learners.

All base-learners in a boosting algorithm must be comparable, so that no base-learner is systematically preferred over others in some iteration of the algorithm. For example, a smooth nonlinear base-learner might fit the negative gradient better than a linear base-learner simply because the nonlinear form has more degrees of freedom rather than because its underlying effects on U are stronger. (Note: smoothness of nonlinear base-learners represented by smoothing or P-splines can be measured by their degrees of freedom, i.e., by trace of their hat matrix, Kneib, Hothorn & Tutz 2009.) In this case, the nonlinear base-learner would be preferred over the linear one, which leads to a biased selection of base-learners. Selecting smoothing parameters to ensure the same degrees of freedom for all base-learners overcomes this problem (see Hastie & Tibshirani 1990; Kneib, Hothorn & Tutz 2009) and controls the smoothness of nonlinear marginal function estimates, so that cross-validation of smoothing parameters for these functions is not needed (unlike conventional GAM estimation, Wood 2006).

COMPONENT-WISE GRADIENT BOOSTING

Component-wise gradient boosting uses only the *one* base-learner that fits U best to estimate U ($\hat{U}$) in each iteration (Bühlmann & Hothorn 2007; see Appendix S1 for a simple example). A small fraction v of $\hat{U}$ is added to the current values $\hat{f}_1, \ldots, \hat{f}_n$ and then $\hat{U}$ is re-estimated in following iterations using the best-fitting base-learner. The step-length factor v needs to be small (see Bühlmann & Hothorn 2007), and we used the commonly chosen value $v = 0{\cdot}1$. The updated values $\hat{f}_1, \ldots, \hat{f}_n$ depend on the predictors because $\hat{U}$ is estimated from a *regression model* (the best-fitting base-learner) that depends on a subset of predictors. If a base-learner has been used in one of the iterations of the gradient boosting algorithm, it is not removed from the set of base-learners but may be used again in later iterations. Selecting the best-fitting base-learner in each step achieves variable selection and helps avoid overfitting.

The component-wise gradient boosting algorithm we used proceeds as follows:

1. Initialize the n-dimensional vector $\hat{f}^{[0]}$ with raw initial estimates of $\hat{f}_1, \ldots, \hat{f}_n$ using zeroes or the average value of the response (Bühlmann & Hothorn 2007).

2. Specify the set of base-learners, as regression estimators with one continuous output variable. Each base-learner depends on one of the p predictors, yielding a set with $P \geq$ p base-learners (because it is possible to specify more than one base-learner for each of the covariates, see below). Set $m = 0$, where m denotes the number of the current boosting iteration.

3. Increase m by 1, derive the negative gradient $-\left(\frac{\partial R}{\partial f_1}, \ldots, \frac{\partial R}{\partial f_n}\right)$ and evaluate it at the current estimates $\hat{f}^{[m-1]} = \left(\hat{f}_1^{[m-1]}, \ldots, \hat{f}_n^{[m-1]}\right)$ to obtain the negative gradient vector $U^{[m-1]} := \left(-\frac{\partial}{\partial f_i}\rho\left(Y_i, \hat{f}_i^{[m-1]}\right)\right)_{i=1,\ldots,n}$.

4. Fit the negative gradient to each base-learner separately to obtain p vectors of predicted values, where each vector is an estimate of the negative gradient vector $U^{[m-1]}$. Select the base-learner that fits $U^{[m-1]}$ best according to the R-squared goodness-of-fit criterion (an appropriate measure of fit because the empirical risk R and its negative gradient $-\left(\frac{\partial R}{\partial f_1}, \ldots, \frac{\partial R}{\partial f_n}\right)$ are both continuous). The vector of predicted values $\hat{U}^{[m-1]}$ from the best-fitting base-learner depends on

the values of predictors that are the inputs of that base-learner.

5. Update $\hat{f}^{[m]} = \hat{f}^{[m-1]} + v\hat{U}^{[m-1]}$ (where v is the step-length) to add a small fraction of the estimated negative gradient to the current values of $\hat{f}_1, \ldots, \hat{f}_n$.

6. Iterate steps 3–5 until a stopping iteration (denoted by m_{stop}) is reached, yielding the final predicted values $\hat{f}^{[m_{\text{stop}}]}$ that estimate the optimal prediction function.

Some base-learners may not be selected before the algorithm stops, achieving automatic variable selection. The algorithm selects among variables and between linear and nonlinear base-learners for some variables, achieving model choice. The estimate of the prediction function has a smaller absolute value than it would if the boosting algorithm had been run until convergence, achieving the shrinkage property of boosting (estimates shrunken towards zero). Variable selection and shrinkage help avoid overfitting so that spurious effects are excluded or diminished. From step 5 it is seen that the final estimates have an additive structure that matches the additive structure of the prediction function f_{GAM} (Friedman, Hastie & Tibshirani 2000; Bühlmann & Hothorn 2007). See Appendix S1 for methodological details.

Gradient boosting is typically stopped early (before convergence) to avoid overfitting the data and improve prediction accuracy. We used fivefold cross-validation to choose the stopping iteration, m_{stop}. In each iteration, we evaluated empirical risk five times using four-fifths of the data to fit the data and the remaining one-fifth to evaluate risk. The five estimates are averaged to obtain the mean empirical risk for that iteration. m_{stop} is the iteration with lowest empirical risk. Because prediction accuracy is optimized within the GAM framework, boosted GAMs can predict novel observations better than GAMs fitted with conventional methods such as backfitting (Hastie & Tibshirani 1990).

SPATIAL AUTOCORRELATION

Spatial autocorrelation is common in ecological data (Legendre 1993). Its presence violates a main assumption of statistical models that rely on independent observations, and failure to account for spatial autocorrelation can bias results and conclusions. Spatial autocorrelation can be addressed by filtering it out prior to modelling, by directly accounting for it in model construction or using permutation (e.g. the Mantel test) to avoid biased significance tests (Legendre 1993; Diggle & Ribeiro 2007). Unlike linear models that represent spatial correlation within the correlation structure of the error term, GLMs and GAMs have no standard formulation to represent spatially autocorrelated count or binary data (Kneib, Müller & Hothorn 2008). Smoothing functions can address spatial autocorrelation in GLMs and GAMs using a smooth, nonlinear, surface function (f_{SP}) of the spatial coordinates (Kneib, Müller & Hothorn 2008). This function, which can be interpreted as the realization of a spatially correlated stochastic process, becomes an additional predictor in the GAM when it is added to the other effects contained in the prediction function. Usually, this function is estimated via two-dimensional spline smoothing of the spatial coordinates (Wood 2003, 2006; Kneib, Müller & Hothorn 2008).

DATA SETS

We studied the 23 408 km^2 portion of Maryland within the Chesapeake Bay basin in the mid-Atlantic region of the United States (online Appendix S2). Benthic macroinvertebrate and fish assemblage data were collected by the Maryland Biological Stream Survey (MBSS, USEPA 1999), an ongoing statewide survey of first- to fourth-order streams where stream physical, hydrological, water chemistry, location, riparian conditions and biological communities are measured (MD DNR 2007). We used data collected from 1994 to 2004 at sites with watershed areas <200 km^2 that were within the Chesapeake Bay watershed. We used only the first record for sites that were sampled more than once ($n = 26$ sites). Of approximately 2500 MBSS samples, 1573 satisfied these conditions.

For macroinvertebrates, we examined taxonomic richness (macroinvertebrate richness); relative abundances of sensitive macroinvertebrates in the Orders Ephemeroptera, Plecoptera and Trichoptera (%EPT); individuals that cling to substrate (%Clingers); and individuals in the collector/gatherer functional feeding group (%Collectors). For fish, we calculated taxonomic richness (fish richness), benthic species richness (benthic fish richness), biomass (fish biomass) and relative abundance of tolerant individuals (%Tolerant Fish).

We assembled a matrix of predictor variables taken from previous studies or that were measured by the MBSS (see Maloney *et al.* 2009 and online Appendix S3). We included data on 15 watershed attributes (e.g. ecoregion, drainage area land use) and 15 stream attributes (e.g. stream water chemistry, width and habitat quality scores; see online Appendix S3). We also included the latitude and longitude of each sampling reach to account for spatial effects.

Seven sites had missing values for one or more predictors, so they were removed before analyses ($n = 1566$). Ninety-five sites had no fish collected and were not used to examine fish responses ($n = 1471$). An additional five sites had no biomass records for fish, leaving $n = 1466$ for this analysis.

STATISTICAL ANALYSES

A component-wise boosted GAM was constructed for each response. For count responses, we used the Poisson distribution (both fish richness measures) or the negative binomial distribution (macroinvertebrate richness) if the Poisson model did not fit because of overdispersion (Hilbe 2007). The link function used was the natural logarithm. For percentage data, we used a Gaussian distribution after arcsine square-root transformation because preliminary models with untransformed Gaussian, log or square-root-transformed Gaussian, or Gamma distributions did not satisfy model assumptions of homoscedasticity and normality of residuals. For fish biomass, we used a Gamma distribution with the logarithmic link function because this variable was positive and highly right-skewed. Predictors with highly right-skewed distributions were $\log_{10}$

transformed before analyses (Appendix S3). Model assumptions were checked using residual plots. Smooth nonlinear functions were modelled using penalized regression splines with a B-spline basis (P-Splines, Wood 2006; Kneib, Hothorn & Tutz 2009).

For each of the $\tilde{p}$ continuous predictor variables, we specified two base-learners: a *linear* base-learner (a linear regression with U as the outcome variable and the predictor as the only input variable) and a *smooth nonlinear* base-learner (a P-spline with U as the outcome variable and the predictor as the only input variable, Kneib, Hothorn & Tutz 2009). For each of $\bar{p}$ categorical predictor variables, we specified one base-learner as a linear model in which U was the outcome variable and dummy-coded variables representing the predictor were the only input variables. To estimate f_{SP}, we specified two base-learners with U as the outcome variable and UTM easting and northing coordinates $\boldsymbol{X}_E$ and $\boldsymbol{X}_N$ as input variables; one base-learner was a *linear* surface function of $\boldsymbol{X}_E$ and $\boldsymbol{X}_N$, and the other was a smooth *nonlinear* function modelled by a tensor product P-spline (Kneib, Müller & Hothorn 2008; Kneib, Hothorn & Tutz 2009). With these specifications, there were $2 \cdot \tilde{p} + \bar{p} + 2$ base-learners in each GAM:

$$f_{\mathrm{GAM}}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p, \boldsymbol{X}_E, \boldsymbol{X}_N) = \beta_0 + \sum_{j=1}^{p} f_j(\boldsymbol{X}_j) + f_{\mathrm{SP}}(\boldsymbol{X}_E, \boldsymbol{X}_N).$$

To assure base-learner comparability, we set the degrees of freedom to 1 for all base-learners by omitting the intercept term and by adding additional penalties to base-learners for categorical predictor variables. We also added a constant base-learner to account for the model intercept (see Hothorn *et al.* 2010a,b for the details of this procedure). For starting values, we used the loss-minimizing constants that maximize the log-likelihood (Bühlmann & Hothorn 2007), such as the mean response value for arcsin-transformed Gaussian models. We could have used zeros, but constants maximizing the log-likelihood shortened running times.

Once GAMs were fitted, we plotted the marginal function estimates for each continuous predictor in each fitted GAM to visualize the strength and form (linear or nonlinear) of dependency patterns between predictors and responses. Marginal estimates can also be used to predict the expected response for any value of the predictors (Appendix S4), and we calculated predicted responses for the 10th, 50th and 90th percentile of % impervious surface cover and watershed area. Bootstrap samples were used to estimate confidence intervals for these predicted responses.

We used bootstrapping to quantify the precision of boosting algorithms. Each of 100 bootstrap samples from the full data set was used as a training data set to which gradient boosting was applied. The GAM estimates and true outcome values of the 100 bootstrap data sets were used to compute bootstrap confidence intervals and medians for the generalized in-sample R^2 (Nagelkerke 1991; also called pseudo R^2, Everitt 2006). We then tested the prediction accuracy of each GAM using bootstrap cross-validation (Harrell 2001). We obtained 100 prediction functions by applying gradient boosting to the 100 bootstrap samples and applied each prediction function to its out-of-bootstrap observations (observations not in the sample). The cross-validation predictions and the measured response values of the 100 cross-validated out-of-bootstrap data sets were used to compute bootstrap confidence intervals and medians for the out-of-bootstrap pseudo R^2, which measures prediction accuracy for novel observations. Although bootstrap analyses were used to estimate confidence intervals, interpretations of effects were based on models built with the full data set.

We also modelled each response using conventional methods for GAM estimation (backfitting in combination with generalized cross-validation using R package mgcv, see Wood 2006). All analyses were conducted with the R system for statistical computing (R Development Core Team 2010). Boosting estimates were obtained using the R add-on package mboost (Hothorn *et al.* 2010a,b). All add-on packages and sample R-code are described in supporting material (online Appendix S5).

Results

MODEL COMPARISON

For fish richness, in-sample and out-of-sample pseudo R^2 values were equivalent between boosted and conventional

Table 1. Median pseudo R^2 values for boosted and conventionally fitted generalized additive models. In-sample pseudo R^2 values measure the fraction of variation in the training data explained by the model (goodness-of-fit). Out-of-bootstrap R^2 values measure the prediction accuracy. The numbers in parentheses are the 10th and 90th percentile values calculated from the 100 bootstrap samples of the full data set

			In-sample		Out-of-bootstrap	
Assemblage	Metric	Distribution	Boosted GAM	Conventional GAM	Boosted GAM	Conventional GAM
Invertebrates	Richness	Negative Binomial	0·43 (0·36, 0·52)	0·45 (0·42, 0·48)	0·22 (0·00, 0·28)	0·21 (0·10, 0·27)
	%EPT	Gaussian	0·59 (0·56, 0·61)	0·64 (0·62, 0·66)	0·47 (0·42, 0·50)	0·36 (0·26, 0·44)
	%Clingers	Gaussian	0·44 (0·42, 0·46)	0·53 (0·51, 0·56)	0·26 (0·23, 0·30)	0·09 (−0·45, 0·17)
	%Collectors	Gaussian	0·28 (0·25, 0·31)	0·41 (0·37, 0·45)	0·06 (0·02, 0·12)	−0·14 (−0·71, 0·00)
Fish	Richness	Poisson	0·97 (0·96, 0·97)	0·97 (0·96, 0·97)	0·94 (0·92, 0·95)	0·94 (0·92, 0·95)
	%Tolerant	Gaussian	0·55 (0·53, 0·58)	0·65 (0·62, 0·67)	0·38 (0·29, 0·43)	0·17 (−0·56, 0·28)
	Benthic richness	Poisson	0·60 (0·58, 0·62)	0·63 (0·61, 0·65)	0·53 (0·49, 0·56)	0·51 (0·45, 0·55)
	Biomass	Gamma	0·41 (0·37, 0·45)	0·52 (0·49, 0·54)	0·24 (0·13, 0·29)	−0·15 (−0·53, 0·14)

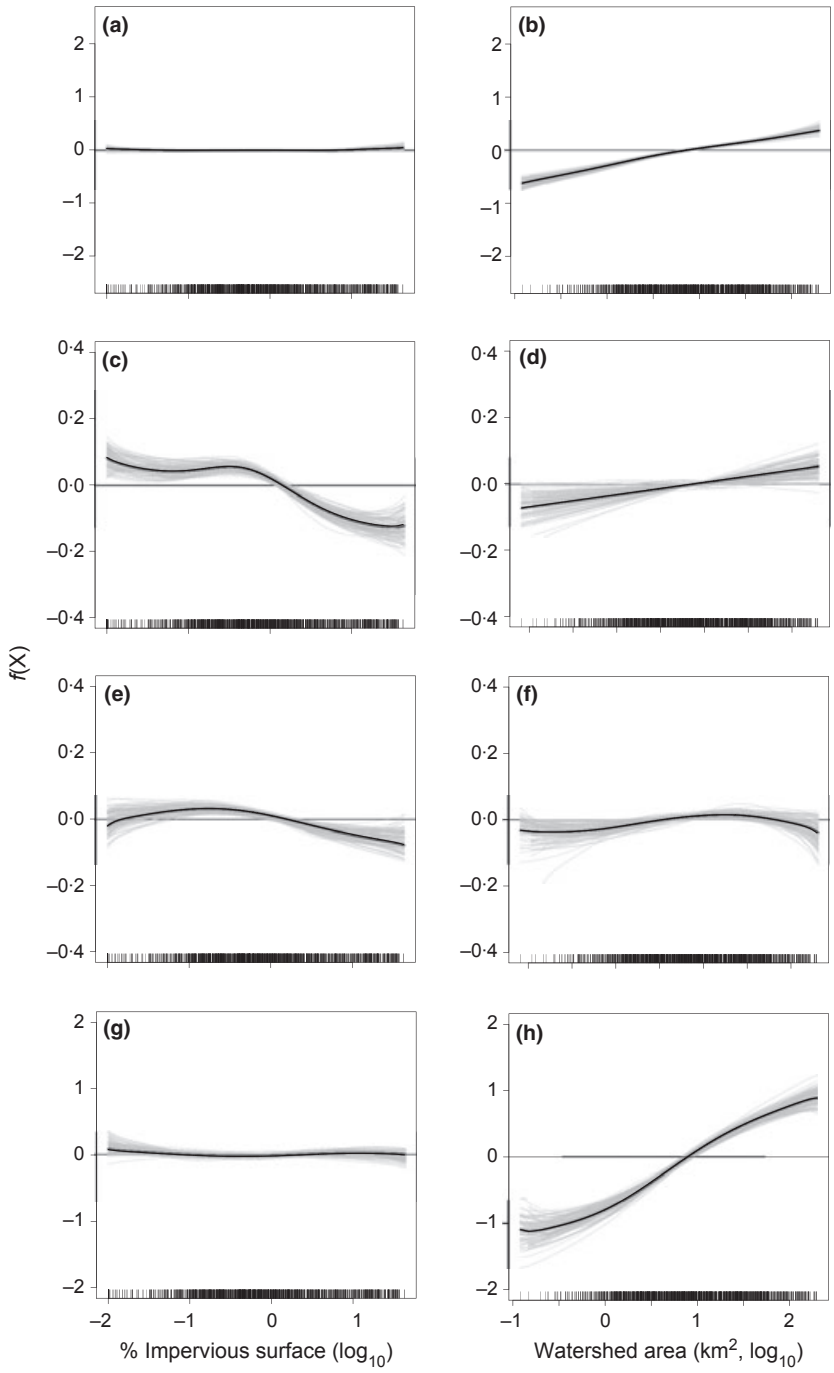

Fig. 2. Marginal functional estimates of boosted generalized additive models for % impervious surface cover in a watershed and watershed area. Response variables were as follows: (a,b) fish richness; (c,d) % of macroinvertebrates collected as Ephemeroptera, Plecoptera and Trichoptera; (e,f) % macroinvertebrates that cling to substrate; and (g,h) number of benthic fish species. Black lines indicate the marginal functional estimates from boosted GAMs using the full data set; grey lines represent marginal functional estimates obtained from 100 bootstrap samples of the full data set. Vertical lines on *x*-axis indicate observed sample values.

GAM models (Table 1). For all other responses, in-sample pseudo R^2 values of conventional GAMs were higher than those of corresponding boosted GAMs; but out-of-bootstrap pseudo R^2 estimates for conventional GAMs were lower than those of boosted GAMs (Table 1) for several responses (%EPT, %Clingers, %Tolerant and Biomass), demonstrating the predictive superiority of these boosted models. We do not present plots of the conventional marginal functions because they are structurally similar to those obtained from the boosted GAMs; however, boosted estimates were less pronounced (closer to zero line) than conventional estimates, probably because of the shrinkage property of gradient boosting. We base the rest of our interpretations on the boosted GAMs.

GOODNESS-OF-FIT OF BOOSTED GAMS

Among responses, the highest amount of variation was explained by the model for fish richness (median in-sample pseudo R^2 = 0·97), and the least variance was explained by the model for %Collectors (median in-sample pseudo R^2 = 0·28; Table 1). Models explained moderate to good amounts of variation for % EPT, % Tolerant Fish and benthic fish richness (median in-sample pseudo R^2 = 0·59, 0·55 and 0·60, respectively). Models for macroinvertebrate richness, %Clingers and fish biomass explained slightly less variation (median in-sample pseudo R^2 = 0·43, 0·44 and 0·41, respectively). Prediction accuracy was the highest for fish richness (out-of-bootstrap median pseudo R^2 = 0·94) and least for %Collectors (out-of-bootstrap median pseudo R^2 = 0·06) and models for other responses fell in between (Table 1). Out-of-bootstrap confidence intervals for predictive pseudo R^2 values indicated that boosted models predicted better than chance alone (Table 1), except for macroinvertebrate richness.

PREDICTOR–RESPONSE RELATIONSHIPS FROM BOOSTED GAMS

Plots of marginal functional estimates show relationships between a response and any predictor variable after accounting for all other covariates. Macroinvertebrate richness had positive and negative relationships with many predictors, including land use (% impervious surface and % pasture), population density, stream water chemistry (pH, conductivity, DOC and NO_3–N), and distance to a main stream or major tributary (online Appendix S6). %EPT was affected by several land uses (% impervious surface and % row crop), natural watershed attributes (elevation and calcareous rock), stream chemistry (DOC) and stream gradient; %Clingers related to elevation, calcareous bedrock and embeddedness and less strongly to % impervious cover than richness and %EPT. %Collectors related to calcareous bedrock and population density. All fish response variables were affected by measures of system size (watershed area and stream width; online Appendix S6). %Tolerant Fish and fish biomass also related to stream water chemistry (pH and DOC) and habitat quality (instream habitat and pool quality). Ecoregion had no apparent effects on any macroinvertebrate or fish response (bootstrap confidence intervals all included 0, Appendix S7).

We present marginal function plots in detail only for % impervious surface cover and watershed area for fish richness, %EPT, %Clingers and number of benthic fish species (the four best models, Fig. 2). There are negative nonlinear effects of impervious surface on %EPT (Fig. 2c), and %Clingers (Fig. 2e). A change in slope appears to occur at *c.* 0·4% for %EPT and *c.* 0·3% for %Clingers. The predicted responses crossed the 0·0 reference line (which represents model-predicted average of the response) at *c.* 1·5% (bootstrap range 1·0–2·0%) for %EPT (Figs 2c) and *c.* 2·0% (range 1·0–3·5%) for %Clingers (Fig. 2e).

Comparing predicted responses at specific values of a predictor can also help quantify predictor–response relationships. With all predictor values set to their respective empirical means, the average prediction of the boosted negative binomial GAM for macroinvertebrate richness was 19·2 (Table 2). This value was included in the back-transformed 90% confidence intervals for the 10th percentile of impervious surface but not for the 50th and 90th percentiles of impervious surface, indicating that watersheds with low amounts of impervious cover (i.e. near the 10th percentile) had model-average macroinvertebrate richness, but watersheds in the 50th and 90th percentiles of impervious cover had richness levels different from the model average (Table 2). Predicted marginal responses for the 10th, 50th and 90th percentiles of impervious surface also showed a decreasing trend (19·8, 19·8 and 17·2, respectively, Table 2). A similar decreasing trend was shown for %EPT and %Clingers (Table 2). The % impervious surface affected both measures of fish richness very weakly or not at all (Fig. 2a,g), which is confirmed by the model-average predictions because each response fell within the 90% confidence interval (Table 2).

The %Clingers was nonlinearly related to watershed area (Fig. 2f), and the effect strength increased up to *c.* 10 km^2, after which it increased more slowly or declined. Marginal function estimates of fish richness and %EPT increased linearly with watershed area (Fig. 2b,d, respectively). Benthic fish richness increased nonlinearly with watershed area (Fig. 2h). Small watersheds (10th percentile) had lower than model-average macroinvertebrate richness, fish richness and benthic fish richness; medium-sized watersheds (50th percentile) had model-average levels of these metrics; and larger watersheds (90th percentile) had higher than model-average levels of these metrics; %Tolerant Fish showed an opposite pattern (Table 2). %EPT and %Clingers did not differ among tested watershed sizes (i.e. all confidence intervals for each tested percentile included the respective average prediction, Table 2).

SPATIAL EFFECTS FROM BOOSTED GAMS

Marginal spatial effects (effects of location after accounting for other covariates) were minor, indicating that most of the observed spatial variation could be explained by the predictor variables. Marginal spatial effects were more apparent in macroinvertebrate than fish responses. We focus on the marginal spatial effects for macroinvertebrate and fish richness (Fig. 3); results for other responses are included in online Appendices S8 and S9. For macroinvertebrate richness, lower values were located in the northern Blue Ridge and Northern Piedmont ecoregions (indicated by darker shading in Fig. 3a). Fish richness was slightly lower in the far western area of the study area (Fig. 3b).

Discussion

We applied gradient boosting to better understand the complex, often nonlinear and spatially correlated effects of anthropogenic activities and natural watershed attributes on stream macroinvertebrates and fish. Boosted GAMs for four responses (%EPT, %Clingers, %Tolerant Fish, and fish biomass) explained a higher proportion of variation in

Table 2. Estimated intercepts and marginal predictions of responses to selected percentiles of % impervious surface cover and watershed area. GAM equations were applied to the marginal predictions at the 10th, 50th and 90th percentiles of % impervious surface and watershed area while other predictors were represented by average values of their marginal prediction functions. Predicted responses were back-transformed to the original measurement scales of responses using the inverse link function *g* (Appendix S4). Numbers in brackets are the indicated percentile values of % impervious surface cover or watershed area. Numbers in parentheses are bootstrapped 90% confidence intervals estimated by the 5% and 95% percentiles from the 100 bootstrap samples. Bold type indicates that the 90% confidence interval does not include the average response, indicating that the response for a predictor percentile is different from modelled average

				% Impervious surface		
Taxa group	Metric	Intercept	Average prediction	10th [0·1%]	50th [0·7%]	90th [14·9%]
Invertebrates	Richness	2·98 (2·94, 2·99)	19·2	19·8 (19·0, 20·5)	**19·8 (19·3, 20·0)**	**17·2 (16·6, 18·7)**
	%EPT	0·55 (0·54, 0·56)	27·3	**31·1 (29·1, 34·1)**	**30·7 (29·2, 32·1)**	**17·7 (15·3, 20·2)**
	%Clingers	0·66 (0·65, 0·67)	37·6	**40·2 (38·5, 42·4)**	**39·3 (38·1, 40·1)**	**32·4 (29·9, 35·0)**
	%Collectors	0·64 (0·63, 0·65)	35·4	35·4 (34·8, 37·0)	35·0 (34·2, 35·5)	36·1 (34·4, 37·1)
				10th [0·1%]	50th [0·7%]	90th [14·5%]
Fish	Richness	2·13 (2·07, 2·27)	8·06	7·97 (7·64, 8·18)	7·99 (7·83, 8·21)	8·30 (7·94, 8·80)
	%Tolerant	0·93 (0·91, 0·94)	63·9	64·1 (62·4, 65·8)	64·3 (63·4, 65·1)	63·3 (61·7, 66·1)
	Biomass	2·15 (2·09, 2·18)	7·38	7·45 (7·38, 8·25)	7·38 (7·18, 7·54)	7·29 (6·68, 7·52)
	Benthic rich.	0·44 (0·34, 0·52)	1·26	1·27 (1·22, 1·33)	1·24 (1·21, 1·28)	1·30 (1·20, 1·36)

		Watershed area (km^2)		
Taxa group	Metric	10th [1·2 km^2]	50th [7·5 km^2]	90th [56·4 km^2]
Invertebrates	Richness	**18·7 (17·5, 19·1)**	19·2 (18·4, 19·3)	**19·6 (19·3, 21·8)**
	%EPT	24·6 (22·9, 27·3)	27·2 (26·7, 27·8)	30·2 (27·3, 32·2)
	%Clingers	35·7 (34·0, 37·6)	38·6 (37·6, 39·3)	37·9 (36·4, 39·9)
	%Collectors	36·4 (35·4, 39·0)	35·4 (34·3, 36·0)	34·4 (32·1, 35·6)
		10th [1·3 km^2]	50th [8·5 km^2]	90th [58·0 km^2]
Fish	Richness	**4·95 (4·58, 5·30)**	8·33 (8·03, 8·54)	**12·44 (11·72, 13·52)**
	%Tolerant	**76·2 (72·5, 79·4)**	63·8 (61·3, 64·7)	**50·7 (48·0, 54·9)**
	Biomass	**5·88 (4·92, 6·68)**	**8·04 (7·64, 8·49)**	8·03 (7·32, 9·29)
	Benthic richness	**0·63 (0·56, 0·71)**	1·31 (1·25, 1·36)	**2·41 (2·20, 2·66)**

out-of-bootstrap samples than conventional GAMs, indicating that gradient boosting improved predictive ability, possibly because conventional GAMs overfit these data. The boosted GAMs also identified many nonlinear relationships between predictors and measures of fish and macroinvertebrate assemblage structure, and identified areas of residual spatial correlation after accounting for other predictors. Measures of system size influenced both fish and macroinvertebrates; however, measures of anthropogenic land use more heavily influenced macroinvertebrates than fish. Boosting improved predictive accuracy compared with conventional GAMs yet preserved the interpretability of conventionally fitted GAMs.

PERFORMANCE AND INTERPRETATION OF BOOSTED GAMS

Fish

Conventional and boosted GAMs for fish richness performed exceptionally and equally well even when predicting out-of-bootstrap samples. Such exceptional fits indicate models with strong inferential and predictive power and suggest that the covariates selected can effectively model fish richness patterns in Maryland. Model results indicate that fish richness in small headwater streams of Maryland is mainly controlled by system size and local habitat factors and it is only minimally affected by catchment land use. The strong, positive effect of system size and local habitat supports previous findings (Schlosser 1982; Angermeier & Schlosser 1989). The weak effect of land use contradicts previous work (e.g. Wang & Lyons 2003; Allan 2004) and demonstrates the difficulty in using fish as indicators of small-stream impairment from anthropogenic activities (Schlosser 1990).

The boosted model for benthic fish richness performed moderately well and slightly better than the conventional GAM in predicting out-of-bootstrap samples. Thus, we have moderate confidence in interpreting the effect estimates. Benthic richness was strongly related to drainage area, further supporting the strong effect of system size on fish assemblages. However, benthic richness also showed a strong effect of human population density, indicating a possible sensitivity of this portion of the fish assemblage to anthropogenic stress.

Boosted GAMs for %Tolerant fish and fish biomass performed moderately well on in-sample data, but poorly with

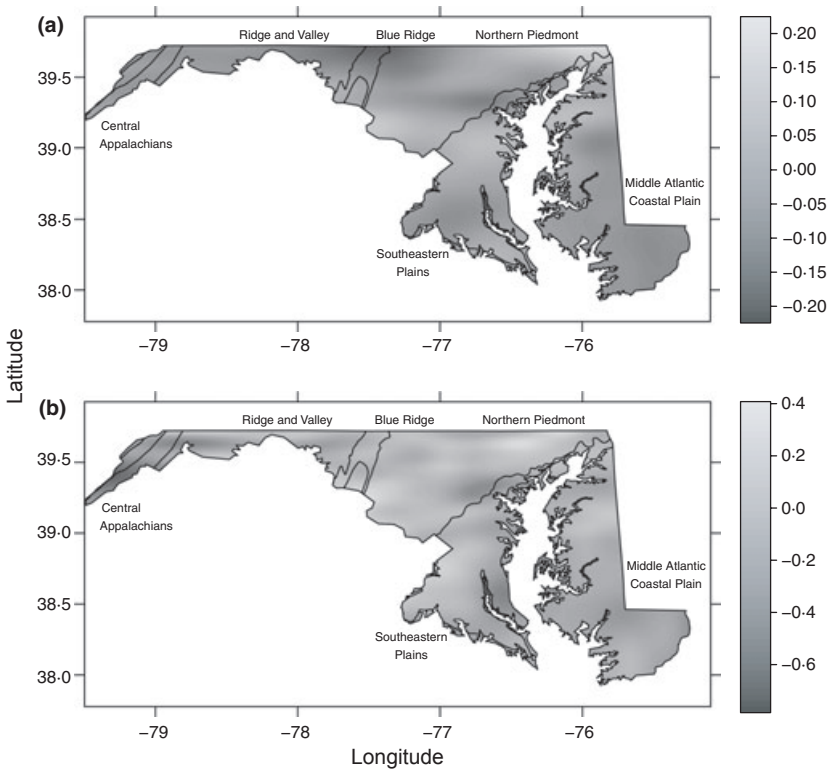

Fig. 3. Marginal spatial effect estimates from the negative binomial model for macroinvertebrate richness (a) and the Poisson model for fish richness (b). Darker shading indicates smaller values of the marginal spatial effect.

out-of-bootstrap data. Thus, we have moderate confidence for inferences within the data set, but lower confidence in predictive ability. Both metrics were strongly related to system size and local habitat. %Tolerant fish was also positively related to human population density, which was expected because this metric represents the fish species tolerant to human stressors.

Benthic macroinvertebrates

The boosted GAM for %EPT performed moderately well for both in-sample and out-of-sample data. The %EPT (percentage of mayflies, stoneflies and caddisflies) represents the macroinvertebrate taxa that are sensitive to anthropogenic stressors, so the strong effects of impervious surface and population density confirm previous studies (e.g. Paul & Meyer 2001; Wang & Lyons 2003). Bootstrapped marginal function estimates (grey lines in Fig. 2) for % impervious surface cover show high confidence that a change in the slope occurred at 0·4% impervious surface, with lower than model-average predictions occurring above 2·0% impervious surface. This level is close to the threshold for negative impacts of impervious surface on macroinvertebrates reported by Baker & King (2010). The low threshold values suggest that the macroinvertebrate assemblage is sensitive to levels of imperviousness far below the 10–20% values in earlier reports (Paul & Meyer 2001; Wang & Lyons 2003).

Boosted GAMs for macroinvertebrate richness and %Clingers both performed moderately well with in-sample data but only weakly in predicting out-of-sample bootstrap data. We have moderate confidence for inferences within the data set but lower confidence in model predictive ability. Marginal function estimates for impervious surface for both boosted GAMs indicate a change in slope at *c*. 1·0% for macroinvertebrate richness and 0·3% for %Clingers, with lower than model-averages occurring at 3·0%, and 1·5% impervious surface, respectively. Both ranges are close to those for %EPT, further supporting a higher sensitivity of these taxa to impervious surface than previously reported. The boosted GAM for %Collectors performed very poorly for both in-sample and out-of-sample data, so we have little to no confidence in estimates from this model.

Spatial effects

Marginal spatial effects from boosted GAMs were more apparent in the macroinvertebrate than fish responses, but were minor (Fig. 3 and online Appendices S8 and S9). Spatial autocorrelation is an important issue in ecological studies (Borcard, Legendre & Drapeau 1992) and biological assessment of streams (King *et al*. 2005). Our relatively small residual spatial effects indicate that our other predictor variables accounted for most of the spatial patterns in the responses, probably because we included predictors that captured important spatial differences (e.g. elevation and bedrock composition) or that were good predictors of responses regardless of spatial patterns. Our study area was irregularly shaped, which

can lead to boundary effects (Fortin 1999), but penalized tensor product B-spline surfaces tend to be relatively robust in cases where data are irregularly distributed (Kneib, Müller & Hothorn 2008).

COLLINEARITY

Some of our predictor variables were highly correlated (Spearman correlation coefficients $>0{\cdot}5$, Appendix S10). In conventionally fitted GAMs, high correlations can lead to multicollinearity problems, such as large variances of estimates and numerical instability in the fitting process. Boosting reduces multicollinearity problems by shrinking effect estimates, which also reduces the variances of estimates. Other popular methods, such as ridge regression, also use shrinkage to address multicollinearity problems and to stabilize model estimation (Hastie, Tibshirani & Friedman 2009). The narrow ranges obtained from the bootstrap analysis (grey lines in Fig. 2 and Appendix S6) confirm the stability of our results by demonstrating that our estimates have small variances. Multicollinearity does not seem to be disrupting our analyses.

IMPROVING MODEL PERFORMANCE

Model performance might be improved by incorporating more effective predictor variables, by refining the statistical methods or by choosing other techniques if prediction (rather than interpretation) is the main goal. More effective measures of the spatial configuration of land uses and quantitative measures of riparian conditions may improve model fits (Baker, Weller & Jordan 2006). Additionally, the MBSS measures of benthic habitat condition were largely qualitative (e.g. % embeddedness), and more quantitative measures of benthic habitat condition, such as bed stability, might improve model performance, especially for responses that reflect benthic conditions (e.g. benthic fish richness and %Clingers). Fish biomass was measured per area rather than volume, so adjusting for volume sampled may improve GAM predictions for this response. The boosted GAM for %Collectors did not fit well. Collectors are the portion of the assemblage that feed on detrital deposits or loose surface films. Additional covariates that quantify these stream characteristics (e.g. benthic particulate organic matter, total suspended solids and a quantitative measure of benthic condition) would likely improve model performance. Spatial patterns were minor for all boosted GAMs, but better measures or models of network position might further dampen the spatial signal.

Emerging statistical refinements may improve the model performance. Alternative distributions (e.g. binomial or beta) or new transformations may better represent percentage data. Boosting might be combined with new statistical techniques, such as GAMs for location, shape and scale (GAMLSS, Rigby & Stasinopoulos 2005), which can model relationships of predictor variables to the scale and shape parameters of a model family as well as to the conditional mean of the response variable. The main effects framework used here could be extended with interaction and varying coefficient terms (Hastie & Tibshirani 1993). All these enhancements might improve fit and predictive ability, but would add a large computational burden, so their feasibilities need to be investigated. A simple test for the importance of interaction would be to compare the predictive ability of boosted GAMs to methods that easily allow interactions (e.g. random forests).

When prediction success is the main goal, tree-based techniques such as bagging or random forests (Breiman 2001) might perform better than boosted GAMs. Because bagging and random forests are tree-based methods, they can easily represent complex interactions between predictor variables that are difficult to incorporate into the additive prediction function of a GAM. However, estimates from bagging and random forests are typically hard to interpret (Cutler *et al.* 2007). Multivariate Adaptive Regression Splines (MARS, Friedman 1991) is another modelling procedure that allows nonlinear predictor–response relationships, automatic variable selection, and detection and inclusion of interactions. MARS was designed for regression problems using the identity link function, so it is less flexible than boosting in handling model families with other link functions or unknown scale parameters (such as contained in the negative binomial family or gamma family). Penalized estimation techniques have also been adapted to the Bayesian framework. For example, Park & Casella (2008) suggested a Bayesian variant of the lasso that accomplishes shrinkage and variable selection.

Repeated sampling of sites or replicated sampling within a site can increase model performance, but is usually impractical for large data sets.

APPLYING BOOSTING TO GAMS IN ECOLOGY

Generalized additive models are widely used in ecology, but it is typically difficult to select a subset of informative covariates and to decide between linear or nonlinear functions for continuous predictor variables. Component-wise gradient boosting simplifies GAM modelling by making these choices automatically, and boosting typically increases predictive ability while preserving the interpretability of marginal predictor effects. With the MBSS example, boosting did not improve variable selection or model choice over traditional GAMs, probably because of strong predictor effects and because the number of sites was relatively large compared with the number of covariates. When there are more predictor variables with weaker effects, boosted GAMs are more likely to outperform conventional methods in variable selection or model choice because conventional methods are prone to overfit with many weak effects. When GAMs are used for prediction, boosting would increase confidence in the predictions. If prediction is not the main focus, it is not necessary to use boosting instead of conventional GAMs to obtain a model with good interpretability. However, if there are many predictor variables in a data set, boosting may still be superior to conventional GAMs because boosting reduces the effects of multicollinearity, overfitting and spurious predictors to improve interpretability and the understanding of functional relationships.

Acknowledgements

We thank the Maryland Department of Natural Resources for providing the MBSS data; Sergej Potapov for providing code to produce the marginal spatial effect plots and Mark Minton, Barbara St John White, Lori Davias, Eric Lind, Jana McPherson and three anonymous reviewers for comments that greatly improved the manuscript. Funding was provided by a Smithsonian Institution Postdoctoral Fellowship awarded to KOM and by the US Environmental Protection Agency National Center for Environmental Research (NCER) Science to Achieve Results (STAR) grant #R831369. The work of MS was supported by the Interdisciplinary Center for Clinical Research (IZKF) at the University Hospital of the University of Erlangen-Nuremberg (Project J11). Use of trade, product or firm names does not imply endorsement by the U.S. Government.

References

Allan, J.D. (2004) Landscapes and riverscapes: the influence of land use on stream ecosystems. *Annual Review of Ecology, Evolution, and Systematics*, **35**, 257–284.

Angermeier, P.L. & Schlosser, I.J. (1989) Species-area relationship for stream fishes. *Ecology*, **70**, 1450–1462.

Austin, M. (2007) Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling*, **200**, 1–19.

Baker, M.E. & King, R.S. (2010) A new method for detecting and interpreting biodiversity and ecological community thresholds. *Methods in Ecology and Evolution*, **1**, 25–37.

Baker, M.E., Weller, D.E. & Jordan, T.E. (2006) Improved methods for quantifying potential nutrient interception by riparian buffers. *Landscape Ecology*, **21**, 1327–1345.

Borcard, D., Legendre, P. & Drapeau, P. (1992) Partialling out the spatial component of ecological variation. *Ecology*, **73**, 1045–1055.

Breiman, L. (1996) Bagging predictors. *Machine Learning*, **24**, 123–140.

Breiman, L. (1998) Arcing classifiers (with discussion). *The Annals of Statistics*, **26**, 801–849.

Breiman, L. (1999) Prediction games and arcing algorithms. *Neural Computation*, **11**, 1493–1517.

Breiman, L. (2001) Random forests. *Machine Learning*, **45**, 5–32.

Bühlmann, P. & Hothorn, T. (2007) Boosting algorithms: regularization, prediction and model fitting (with discussion). *Statistical Science*, **22**, 477–522.

Bühlmann, P. & Yu, B. (2003) Boosting with the L2 loss: regression and classification. *Journal of the American Statistical Association*, **98**, 324–338.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A. & Hess, K.T. (2007) Random forests for classification in ecology. *Ecology*, **88**, 2783–2792.

De'ath, G. (2007) Boosted trees for ecological modeling and prediction. *Ecology*, **88**, 243–251.

Diggle, P.J. & Ribeiro, P.J. (2007) *Model-based Geostatistics*. Springer, New York.

Elith, J., Leathwick, J.R. & Hastie, T. (2008) A working guide to boosted regression trees. *Journal of Animal Ecology*, **77**, 802–813.

Everitt, B.S. (2006) *The Cambridge Dictionary of Statistics*, 3rd edn. Cambridge University Press, Cambridge.

Fortin, M.J. 1999. Spatial statistics in landscape ecology. *Landscape Ecological Analysis: Issues and Applications* (eds J.M. Klopatek & R.H. Gardner), pp. 253–279. Springer, New York.

Freund, Y. & Schapire, R. 1996. Experiments with a New Boosting Algorithm. *Proceedings of the Thirteenth International Conference on Machine Learning Theory*. Morgan Kaufmann Publishers Inc, San Francisco.

Freund, Y. & Schapire, R. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, **55**, 119–139.

Friedman, J.H. (1991) Multivariate Adaptive Regression Splines (with discussion). *Annals of Statistics*, **19**, 1–141.

Friedman, J.H. (2001) Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, **29**, 1189–1232.

Friedman, J.H., Hastie, T. & Tibshirani, R. (2000) Additive logistic regression: a statistical view of boosting (with discussion). *The Annals of Statistics*, **28**, 337–407.

Harrell, F.E. (2001) *Regression Modeling Strategies*. Springer, New York.

Hastie, T. & Tibshirani, R. (1990) *Generalized Additive Models*. Chapman & Hall, London.

Hastie, T.J. & Tibshirani, R.J. (1993) Varying-coefficient models. *Journal of the Royal Statistical Society, Series B*, **55**, 757–796.

Hastie, T., Tibshirani, R. & Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. Springer, New York.

Hilbe, J.M. (2007) *Negative Binomial Regression*. Cambridge University Press, Cambridge.

Hoerl, A. & Kennard, R. (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M. & Hofner, B. (2010a): *mboost: Model-Based Boosting*. R package version 2.0-6. https://r-forge.r-project.org/projects/mboost/.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M. & Hofner, B. (2010b) Model-based boosting 2.0. *Journal of Machine Learning Research*, **11**, 2109–2113.

King, R.S., Baker, M.E., Whigham, D.F., Weller, D.E., Jordan, T.E., Kazyak, P.F. & Hurd, M.K. (2005) Spatial considerations for linking watershed land cover to ecological indicators in streams. *Ecological Applications*, **15**, 137–153.

Kneib, T., Hothorn, T. & Tutz, G. (2009) Variable selection and model choice in geoadditive regression models. *Biometrics*, **65**, 626–634.

Kneib, T., Müller, J. & Hothorn, T. (2008) Spatial smoothing techniques for the assessment of habitat suitability. *Environmental and Ecological Statistics*, **15**, 343–364.

Legendre, P. (1993) Spatial autocorrelation: trouble or new paradigm? *Ecology*, **74**, 1659–1673.

Maloney, K.O., Weller, D.E., Russell, M.J. & Hothorn, T. (2009) Classifying the biological condition of small streams: an example using benthic macroinvertebrates. *Journal of the North American Benthological Society*, **28**, 869–884.

McCullagh, P. & Nelder, J. (1989) *Generalized Linear Models*, 2nd edn. Chapman & Hall/CRC, New York.

MD DNR. 2007. *Maryland Biological Stream Survey, Sampling Manual: Field Protocols*. Maryland Department of Natural Resources, Monitoring and Nontidal Assessment Division, Annapolis, Maryland, USA.

Meier, L., van de Geer, S. & Bühlmann, P. (2009) High-dimensional additive modeling. *The Annals of Statistics*, **37**, 3779–3821.

Nagelkerke, N.J.D. (1991) A note on a general definition of the coefficient of determination. *Biometrika*, **78**, 691–692.

O'Hara, R.B. & Sillanpää, M.J. (2009) A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*, **4**, 85–118.

Park, T. & Casella, G. (2008) The Bayesian Lasso. *Journal of the American Statistical Association*, **103**, 681–686.

Paul, M.J. & Meyer, J.L. (2001) Streams in the urban landscape. *Annual Review of Ecology and Systematics*, **32**, 333–365.

R Development Core Team. 2010. *R: A Language and Environment for Statistical Computing*, v. 2.10.1. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org [accessed 30 March 2010].

Rigby, R.A. & Stasinopoulos, D.M. (2005) Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society, Series C*, **54**, 507–554.

Schlosser, I.J. (1982) Fish community structure and function along two habitat gradients in a headwater stream. *Ecological Monographs*, **52**, 395–414.

Schlosser, I.J. (1990) Environmental variation, life-history attributes, and community structure in stream fishes - implications for environmental management and assessment. *Environmental Management*, **14**, 621–628.

Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.

USEPA. 1999. *From the Mountains to the Sea: The State of Maryland's Freshwater Streams*. EPA/903/R-99/023, Office of Research and Development, United States Environmental Protection Agency, Washington, DC.

Wang, L. & Lyons, J. (2003) Fish and benthic macroinvertebrate assemblages as indicators of stream degradation in urbanizing watersheds. *Biological Response Signatures: Indicator Patterns Using Aquatic Communities* (ed. T.P. Simon), pp. 227–249. CRC Press, New York.

Wood, S.N. (2003) Thin plate regression splines. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **65**, 95–114.

Wood, S.N. (2006) *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, Boca Raton.

Yuan, L.L. & Norton, S.B. (2003) Comparing responses of macroinvertebrate metrics to increasing stress. *Journal of the North American Benthological Society*, **22**, 308–322.

Received 29 June 2010; accepted 28 April 2011
Handling Editor: Jana McPherson

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Detailed description how gradient boosting fits a GAM and a simple example of component-wise gradient boosting.

Appendix S2. Map of the Chesapeake Bay basin portion of Maryland, USA, study locations (*X*), and ecoregion boundaries. Inset shows the position of the study area in relation to the mid-Atlantic region of the USA.

Appendix S3. List of covariates used in model construction. *Variables that were $\log_{10}$ transformed prior to analyses.

Appendix S4. Mathematical formulas for transforming marginal function estimates from boosted GAMs back to the original units of the responses.

Appendix S5. Software and sample R-code used in statistical analyses.

Appendix S6. Marginal functional estimates for boosted generalized additive models for all eight response variables. For figures: black lines indicate the marginal functional estimates from the boosted GAMs using the full data set; gray lines represent marginal functional estimates obtained from 100 bootstrap samples of the full data set. Vertical lines on the *x*-axis indicate observed sample values. Abbreviations for predictors (*x*-axis) defined in Appendix S3.

Appendix S7. Marginal ecoregion effect estimates for boosted generalized additive models for benthic macroinvertebrate and fish response variables. All effects are mean-centered.

Appendix S8. Marginal spatial effect estimates for boosted generalized additive models for benthic macroinvertebrate response variables.

Appendix S9. Marginal spatial effect estimates for boosted generalized additive models for fish response variables.

Appendix S10. Correlation (Spearman) matrix of covariates.

Appendix S11. Fish diversity data file.

Appendix S12. Covariate data file.

As a service to our authors and readers, this journal provides support ing information supplied by the authors. Such materials may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.